Model selection in high-dimensional quantile regression with seamless L_0 penalty

Gabriela CIUPERCA

Université de Lyon, Université Lyon 1, CNRS, UMR 5208, Institut Camille Jordan, Bat. Braconnier, 43, blvd du 11 novembre 1918, F - 69622 Villeurbanne Cedex, France,

Abstract

In this paper we are interested in parameters estimation of linear model when number of parameters increases with sample size. Without any assumption about moments of the model error, we propose and study the seamless L_0 quantile estimator. For this estimator we first give the convergence rate. Afterwards, we prove that it correctly distinguishes between zero and nonzero parameters and that the estimators of the nonzero parameters are asymptotically normal. A consistent BIC criterion to select the tuning parameters is given.

1 Introduction

Consider a model where the number of regressors can increase with the sample size n:

$$Y_i = \mathbf{X}_i^t \boldsymbol{\beta}_n + \varepsilon_i, \qquad i = 1, \cdots, n, \tag{1}$$

where $\boldsymbol{\beta}_n = (\beta_1, \cdots, \beta_{d_n}) \in \mathbb{R}^{d_n}$ contains the regression parameters. The design vector $\mathbf{X}_i$, for observation i, is a deterministic vector of dimension $d_n \times 1$. The random variable ε_i is the model error. Denote by $\boldsymbol{\beta}_n^0 = (\beta_1^0, \cdots, \beta_{d_n}^0)$ the true value, unknown, of the parameter $\boldsymbol{\beta}_n$. In order to automatically select the non-zero components of $\boldsymbol{\beta}_n$ (therefore to select the significant variables), intuitively, the random optimization process would penalize with the "norm" L_0 (it is not a norm) defined by $\|\boldsymbol{\beta}_n\|_0 = \sum_{j=1}^{d_n} \mathbb{1}_{\beta_j \neq 0}$. This "norm" has the disadvantage that it is not continuous in 0, then it is computationally infeasible, since all possible models should be considered (all possible combinations of $\beta_j \neq 0$). In this paper, we estimate the parameter $\boldsymbol{\beta}_n$ of (1), penalizing the quantile process with a seamless L_0 norm. The difficulty in studing of this type of estimation method is that the quantile process is convex in $\boldsymbol{\beta}_n$ and the seamless L_0 penalty is concave.

In literature on the high-dimension models, it was considered only the case of a quantile process penalized with a convex penalty of type L_1. Models with the number of variables exceeding the sample size $(d_n > n)$ are studied by [Belloni and Chernozhukov (2011)], [Fan et al. (2014a)], [Zheng et al.(2013)]. If $d_n < n$, references [Wu and Liu (2009)], [Zou and Yuan(2008)] considered variable selection in a quantile model with convex penalties. Penalized random process of type:

$$G_n(\boldsymbol{\beta}_n) + \mathcal{P}en(\boldsymbol{\beta}_n), \tag{2}$$

with the process $G_n(\boldsymbol{\beta}_n)$ convex in $\boldsymbol{\beta}_n$ and the penalty $\mathcal{P}en(\boldsymbol{\beta}_n)$ nonconvex has been few studied. In [Fan and Peng (2004)], $G_n(\boldsymbol{\beta})$ is $-$loglikelihood and the penalty is nonconvex, with $d_n^5/n \to \infty$, as $n \to \infty$. For $d_n \gg n$, [Wang et al. (2014)] considered, for the particular case of $Y|\mathbf{X} = \mathbf{x}$ sub-Gaussian, $G_n(\boldsymbol{\beta}_n)$ a loss function and $\mathcal{P}en(\boldsymbol{\beta}_n)$ nonconvex loss penalty. For always $d_n \gg n$, [Zhang and Zhang (2012)] considered $G_n(\boldsymbol{\beta}_n) = (2n)^{-1} \sum_{i=1}^n (Y_i - \mathbf{X}_i^t \boldsymbol{\beta}_n)^2$ and $\mathcal{P}en(\boldsymbol{\beta}_n)$ concave. [Fan et al. (2014b)] proposed an estimation method based on one-step local linear approximation, when the support set for $\boldsymbol{\beta}_n^0$ is known.

To overcome the disadvantage of the discontinuity in 0 of the norm L_0, [Dicker et al. (2013)] propose a seamless L_0 penalty:

$$\mathcal{P}en(\boldsymbol{\beta}_n) \equiv \frac{\lambda_n}{\log 2} \sum_{j=1}^{d_n} \log \Big(\frac{|\beta_j|}{|\beta_j| + \gamma_n} + 1 \Big), \tag{3}$$

with $\lambda_n, \gamma_n > 0$ two tuning parameters. If $\gamma_n \to 0$, the penalty L_0 is obtained. Reference [Dicker et al. (2013)] considers $G_n(\boldsymbol{\beta}_n) = n^{-1} \sum_{i=1}^n (Y_i - \mathbf{X}_i^t \boldsymbol{\beta}_n)^2$, with (ε_i) i.i.d., $I\!\!E[\varepsilon_i] = 0$, $Var(\varepsilon_i) = \sigma^2$, suppositions under which the sparsity and the asymptotic normality of estimators are proved, if $d_n/n \to 0$, for $n \to \infty$. If Y belongs to the exponential family, [Li et al. (2012)] considers $G_n(\boldsymbol{\beta}_n) = -\text{loglikelihood}/n$, with penalty (3), but with a stronger constraint on d_n: $d_n^5/n \to 0$ for $n \to \infty$.

If the law of the error ε is unknown, or if the assumptions on the first two moments of the error are not satisfied, then the likelihood, least squares methods with seamless L_0 penalty can not be used. This justifies the interest of the present paper, where quantile process will be penalized with seamless L_0 penalty (3).

We give some general notations. Throughout the paper, C denotes a positives generic constant not dependent on n which may take different values in different formula or even in different parts of the same formula. All vectors and matrices are in bold and all vectors are column. For a vector $\mathbf{v}$, $\|\mathbf{v}\|_2$ is the Euclidean norm, $\mathbf{v}^t$ denotes the transposed of $\mathbf{v}$. For a matrix $\mathbf{M}$, $\|\mathbf{M}\|_2$ is the subordinate norm to the vector norm $\|.\|_2$, $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ are smallest and largest eigenvalues. We use also the notation $sgn(.)$ for the sign function and $tr(.)$ for the trace operator.

The paper is organized as follows. In Section 2, we introduce and study the convergence rate, oracle properties of the seamless L_0 quantile estimator. In Section 3 we propose a consistent BIC criterion to select the tuning parameters. Finally, in Section 3, we present two lemmas useful to prove the main results.

2 Seamless L_0 quantile estimator

In this section we propose and study the seamless L_0 quantile estimator. For a fixed quantile index $\tau \in (0, 1)$, the seamless L_0 quantile estimator is the parameter which minimizes the process

$$Q_n(\boldsymbol{\beta}_n) \equiv \frac{1}{2n} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_i^t \boldsymbol{\beta}_n) + \sum_{j=1}^{d_n} p_{SELO}(\beta_j),$$

with the function $\rho_\tau(.) : \mathbb{R} \to \mathbb{R}_+$ defined by $\rho_\tau(u) = u(\tau - \mathbb{1}_{u<0})$ and for $\beta \in \mathbb{R}$,

$$p_{SELO}(\beta) \equiv \frac{\lambda_n}{\log 2} \log\Big(\frac{|\beta|}{|\beta| + \gamma_n} + 1\Big).$$

Then, the seamless L_0 quantile estimator is

$$\widehat{\boldsymbol{\beta}}_n \equiv \underset{\boldsymbol{\beta}_n \in \mathbb{R}^{d_n}}{\arg\min} \, Q_n(\boldsymbol{\beta}_n). \tag{4}$$

Remark 1 *We emphasize that the results of [Fan et al. (2014b)], where a concave penalty is considered for quantile process, cannot be applied in the present paper, because our penalty cannot written as $\|\mathbf{c} \circ \boldsymbol{\beta}\|_1$.*

For errors (ε_i) of model (1), we consider the following assumption:
(A1) $(\varepsilon_i)_{1 \leq i \leq n}$ are i.i.d., with the distribution function F and density function f. The density function f is continuously, strictly positive in a neighborhood of zero and has a bounded first derivative in the neighborhood of 0. The τth quantile of ε_i is zero: $\tau = F(0)$.
Let us denote $\alpha_n = (d_n/n)^{1/2}$. For the deterministic design $(\mathbf{X}_i)_{1 \leq i \leq n}$ we suppose that:
(A2) there exist constants $0 < r_0 \leq R_0 < \infty$ such that $r_0 \leq \lambda_{\min}(n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^t) \leq \lambda_{\max}(n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^t) \leq R_0$.
(A3) $\max_{1 \leq i \leq n} \|\mathbf{X}_i\|_2 = o(\alpha_n^{-1})$.
On the tuning parameters λ_n, γ_n and on the dimension d_n, we suppose:
(A4) d_n is such that $d_n/n \to 0$, as $n \to \infty$.
(A5) $\lambda_n = O(1)$, $\lambda_n \sqrt{n/d_n} \to \infty$ and $\gamma_n = O(d_n^{1/2} n^{-3/2})$.

Assumptions (A1), (A2) are standard for linear model and (A3) is classic for an high-dimensional model. Assumptions (A4), (A5) are needed for statistical inference study of $\widehat{\boldsymbol{\beta}}_n$ (see e.g. [Dicker et al. (2013)], [Lee et al. (2014)]).

For $\boldsymbol{\beta}_n \in \mathbb{R}^{d_n}$, let be the difference between two quantile processes:

$$G_n(\boldsymbol{\beta}_n) \equiv \sum_{i=1}^n [\rho_\tau(Y_i - \mathbf{X}_i^t \boldsymbol{\beta}_n) - \rho_\tau(\varepsilon_i)]. \tag{5}$$

Following theorem states that the estimators $\widehat{\boldsymbol{\beta}}_n$ has a convergence rate of order α_n. If d_n is bounded, we find the classic convergence rate $n^{-1/2}$ of quantile estimator for a finite-dimensional model (see [Knight (1998)]).

Theorem 1 *Under assumptions (A1)-(A5), we have:* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0\|_2 = O_\mathbb{P}(\alpha_n)$.

Proof. In order to prove the theorem, we show that for all $\epsilon \in (0,1)$, there exists a constant large enough $B > 0$, such that we have, for any n large enough,

$$\mathbb{P}\left[\inf_{\|\mathbf{u}\|_2 = 1} Q_n(\boldsymbol{\beta}_n^0 + B\alpha_n\mathbf{u}) > Q_n(\boldsymbol{\beta}_n^0)\right] \geq 1 - \epsilon. \tag{6}$$

Fix $\epsilon \in (0,1)$ and consider some $\mathbf{u} = (u_1, \cdots, u_{d_n}) \in \mathbb{R}^{d_n}$, with $\|\mathbf{u}\|_2 = 1$. Let be some constant $c > 0$. Consider

$$Q_n(\boldsymbol{\beta}_n^0 + c\alpha_n\mathbf{u}) - Q_n(\boldsymbol{\beta}_n^0) = \frac{1}{2n}G_n(c\alpha_n\mathbf{u}) + \sum_{j=1}^{d_n}[p_{SELO}(\beta_j^0 + c\alpha_n u_j) - p_{SELO}(\beta_j^0)]. \tag{7}$$

For the penalty, we have the following inequality:

$$\sum_{j=1}^{d_n}[p_{SELO}(\beta_j^0 + c\alpha_n u_j) - p_{SELO}(\beta_j^0)] \geq \sum_{j \in J(\mathbf{u})}[p_{SELO}(\beta_j^0 + c\alpha_n u_j) - p_{SELO}(\beta_j^0)],$$

where $J(\mathbf{u}) \equiv \{l \in \{1, \cdots, d_n\};\ p_{SELO}(\beta_l^0 + c\alpha_n u_l) - p_{SELO}(\beta_l^0) < 0\}$. Because c, $\mathbf{u}$ are fixed and $\alpha_n \to 0$, then by Lemma 1, for all $j \in J(\mathbf{u})$, and for large enough n, there exists $\tilde{C}_j > 0$ such that

$$p_{SELO}(\beta_j^0 + c\alpha_n u_j) - p_{SELO}(\beta_j^0) = \frac{\lambda_n}{\log 2}\left[g(\beta_j^0 + c\alpha_n u_j) - g(\beta_j^0)\right] = \frac{\lambda_n}{\log 2}\tilde{C}_j\alpha_n|u_j|\gamma_n(-1)^{sgn(\beta_j^0(\beta_j^0 + c\alpha_n u_j))}.$$

Thus, by assumptions (A4) and (A5), we have:

$$\sum_{j \in J(\mathbf{u})}[p_{SELO}(\beta_j^0 + c\alpha_n u_j) - p_{SELO}(\beta_j^0)] > -\frac{\lambda_n\alpha_n\gamma_n}{\log 2}\sum_{j \in J(\mathbf{u})}\tilde{C}_j|u_j| = -O(\lambda_n\alpha_n\gamma_n d_n) = -O(\alpha_n\alpha_n^{3/2}) = -o(\alpha_n^2). \tag{8}$$

We now study the expectation of $G_n(c\alpha_n\mathbf{u})$:

$$\mathbb{E}[G_n(c\alpha_n\mathbf{u})] = \sum_{i=1}^n \mathbb{E}[\rho_\tau(\varepsilon_i - c\alpha_n\mathbf{X}_i^t\mathbf{u}) - \rho_\tau(\varepsilon_i)] = \sum_{i=1}^n \mathbb{E}\left[\int_0^{c\alpha_n\mathbf{X}_i^t\mathbf{u}} \mathbb{1}_{0 < \varepsilon_i < t}dt\right] = \sum_{i=1}^n \int_0^{c\alpha_n\mathbf{X}_i^t\mathbf{u}}[F(t) - F(0)]dt.$$

On the other hand, by (A1), for $v \to 0$, we have $\int_0^v[F(t) - F(0)]dt = \frac{f(0)}{2}v^2 + o(v^2)$. Using (A3), we have:

$$\frac{1}{n}\sum_{i=1}^n \int_0^{c\alpha_n\mathbf{X}_i^t\mathbf{u}}[F(t) - F(0)]dt = \frac{f(0)}{2}c\alpha_n^2\frac{1}{n}\sum_{i=1}^n(\mathbf{X}_i^t\mathbf{u})^2 + o(\alpha_n^2\frac{1}{n}\sum_{i=1}^n \mathbf{u}^t(\mathbf{X}_i\mathbf{X}_i^t)\mathbf{u}).$$

Then

$$\frac{1}{n}\mathbb{E}[G_n(c\alpha_n\mathbf{u})] = c\frac{f(0)}{2}\alpha_n^2\frac{1}{n}\mathbf{u}^t(\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^t)\mathbf{u}(1 + o(1)). \tag{9}$$

Consider now the random variables $\mathcal{D}_i \equiv (1 - \tau)\mathbb{1}_{\varepsilon_i > 0} - \tau\mathbb{1}_{\varepsilon_i > 0}$, $R_i \equiv \rho_\tau(\varepsilon_i - c\alpha_n\mathbf{X}_i^t\mathbf{u}) - \rho_\tau(\varepsilon_i) - c\alpha_n\mathcal{D}_i\mathbf{X}_i^t\mathbf{u}$ and the random vector $\mathbf{W}_n \equiv \sum_{i=1}^n c\alpha_n\mathcal{D}_i\mathbf{X}_i^t$. Thus, the process G_n can be written:

$$G_n(c\alpha_n\mathbf{u}) = \mathbb{E}[G_n(c\alpha_n\mathbf{u})] + \mathbf{W}_n\mathbf{u} + \sum_{i=1}^n\left[R_i - \mathbb{E}[R_i]\right].$$

But, since, by (A1), the errors (ε_i) are independent, using also $|R_i| \leq |c\alpha_n\mathbf{X}_i^t\mathbf{u}|\mathbb{1}_{|\varepsilon_i| \leq |c\alpha_n\mathbf{X}_i^t\mathbf{u}|}$, we have:

$$\mathbb{E}[\sum_{i=1}^n[R_i - \mathbb{E}[R_i]]^2 = \sum_{i=1}^n \mathbb{E}[R_i - \mathbb{E}[R_i]]^2 \leq \sum_{i=1}^n \mathbb{E}[R_i]^2 \leq \sum_{i=1}^n |c\alpha_n\mathbf{X}_i^t\mathbf{u}|^2\mathbb{E}[\mathbb{1}_{|\varepsilon_i| \leq |c\alpha_n\mathbf{X}_i^t\mathbf{u}|}].$$

3

Taking into account assumptions (A1) and (A3), we have $\mathbb{E}[\mathbb{1}_{|\varepsilon_i| \leq |c\alpha_n \mathbf{X}_i^t \mathbf{u}|}] = F(|c\alpha_n \mathbf{X}_i^t \mathbf{u}|) - F(-|c\alpha_n \mathbf{X}_i^t \mathbf{u}|) = C\alpha_n |\mathbf{X}_i^t \mathbf{u}| \leq C\alpha_n \max_{1 \leq i \leq n} \|\mathbf{X}_i\|_2 = o(1)$, with $C > 0$. Then, using assumption (A2), we obtain:

$$\mathbb{E}\Big[\sum_{i=1}^n [R_i - \mathbb{E}[R_i]]\Big]^2 = o\Big(\alpha_n^2 \mathbf{u}^t \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^t \mathbf{u}\Big) = o(d_n). \tag{10}$$

Consider now the random variable $U_n \equiv d_n^{-1/2} \sum_{i=1}^n [R_i - \mathbb{E}[R_i]]$. Taking into account (10), we have $\mathbb{E}[U_n^2] = o(1)$. Since $E[U_n] = 0$, by Bienaymé-Tchebychev inequality, we have $U_n \xrightarrow[n \to \infty]{\mathbb{P}} 0$. Thus $\sum_{i=1}^n [R_i - \mathbb{E}[R_i]] = o_{\mathbb{P}}(d_n^{1/2})$. Returning to G_n, we have, taking into account (9):

$$G_n(c\alpha_n \mathbf{u}) = \mathbb{E}[G_n(c\alpha_n \mathbf{u})] + \mathbf{W}_n \mathbf{u} + o_{\mathbb{P}}(d_n^{1/2})$$

or again

$$G_n(c\alpha_n \mathbf{u}) = \left(\frac{f(0)}{2} c^2 d_n \mathbf{u}^t \Big(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^t\Big)\mathbf{u} + d_n^{1/2} c \Big(\sum_{i=1}^n \frac{\mathcal{D}_i \mathbf{X}_i^t}{\sqrt{n}}\Big)\mathbf{u}\right)(1 + o_{\mathbb{P}}(1)) + o_{\mathbb{P}}(d_n^{1/2}). \tag{11}$$

Since $n^{-1/2} \sum_{i=1}^n \mathcal{D}_i \mathbf{X}_i^t \mathbf{u}$ converges in distribution to a centered normal distribution, by assumptions (A4) and (A2), for a large enough constant B, we have that the first term of the right side that will dominate in (11). Then,

$$\frac{1}{2n} G_n(B\alpha_n \mathbf{u}) = f(0) B^2 \alpha_n^2 \mathbf{u}^t \Big(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^t\Big)\mathbf{u}(1 + o_{\mathbb{P}}(1)). \tag{12}$$

Thus, for n and B large enough, we have $(2n)^{-1} G_n(B\alpha_n \mathbf{u}) > 0$. On the other hand, by relations (7) and (8), $Q_n(\boldsymbol{\beta}_n^0 + B\alpha_n \mathbf{u}) - Q_n(\boldsymbol{\beta}_n^0) > (2n)^{-1} G_n(B\alpha_n \mathbf{u}) - o(\alpha_n^2)$. Taking also into account relation (12) and assumption (A2), we obtain (6). $\blacksquare$

Let us consider the parameter set, with the constant $B > 0$ of relation (6):

$$\mathcal{V}_{\alpha_n}(\boldsymbol{\beta}_n^0) \equiv \{\boldsymbol{\beta}_n \in \mathbb{R}^{d_n}; \quad \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\| \leq B\alpha_n\}.$$

According to Theorem 1, the seamless L_0 quantile estimators belong to $\mathcal{V}_{\alpha_n}(\boldsymbol{\beta}_n^0)$, with a probability converging to 1. For the index set $\mathcal{A}$, with $\mathcal{A} \subseteq \{1, \cdots, d_n\}$, we will denote by $|\mathcal{A}|$ its cardinal. Throughout the paper, we denote by $\boldsymbol{\beta}_{\mathcal{A}}$ the sub-vector of $\boldsymbol{\beta}_n$ containing the corresponding components of $\mathcal{A}$. Similarly for $\mathbf{X}_{i,\mathcal{A}}$. Consider also the following index set:

$$\mathcal{A}^0 \equiv \{j \in \{1, \cdots, d_n\}; \quad \beta_j^0 \neq 0\}. \tag{13}$$

The following theorem gives the oracle properties for the estimators $\widehat{\boldsymbol{\beta}}_n = (\widehat{\beta}_1, \cdots, \widehat{\beta}_{d_n})$, defined by (4). Note that, with respect to the paper of [Dicker et al. (2013)], for showing the normality of the nonzero estimators, the condition $\mathbb{E}[|\varepsilon_i|^{2+\delta}] < M$ is not needed, for some $\delta > 0$ and $M < \infty$.

Theorem 2 *Under assumptions (A1)-(A5), we have*
(i) $\lim_{n \to \infty} \mathbb{P}[\{j \in \{1, \cdots, d_n\}; \widehat{\beta}_j \neq 0\} = \mathcal{A}^0] = 1$.
(ii) For any vector $\mathbf{u}$ of dimension $|\mathcal{A}^0|$ such that $\|\mathbf{u}\|_2 = 1$, if we denote $\boldsymbol{\Sigma}_{\mathcal{A}^0} \equiv n^{-1} \sum_{i=1}^n \mathbf{X}_{i,\mathcal{A}^0} \mathbf{X}_{i,\mathcal{A}^0}^t$, then

$$\sqrt{n}(\mathbf{u}^t \boldsymbol{\Sigma}_{\mathcal{A}^0}^{-1} \mathbf{u})^{-1/2} \mathbf{u}^t (\widehat{\boldsymbol{\beta}}_{\mathcal{A}^0} - \boldsymbol{\beta}_{\mathcal{A}^0}^0) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}\Big(0, \frac{\tau(1-\tau)}{f^2(0)}\Big).$$

Proof. *(i)* If we denote by $\mathcal{A}^{0c}$ the complementary set of $\mathcal{A}^0$ in $\{1, \cdots, d_n\}$, we will prove that for any $\boldsymbol{\beta}_n = (\boldsymbol{\beta}_{\mathcal{A}^0}, \boldsymbol{\beta}_{\mathcal{A}^{0c}}) \in \mathcal{V}_{\alpha_n}(\boldsymbol{\beta}_n^0)$ such that $\|\boldsymbol{\beta}_{\mathcal{A}^0} - \boldsymbol{\beta}_{\mathcal{A}^0}^0\|_2 = O_{\mathbb{P}}(\alpha_n)$ and any constant $C > 0$, we have

$$Q_n((\boldsymbol{\beta}_{\mathcal{A}^0}, \mathbf{0})) = \min_{\|\boldsymbol{\beta}_{\mathcal{A}^{0c}}\| \leq C\alpha_n} Q_n((\boldsymbol{\beta}_{\mathcal{A}^0}, \boldsymbol{\beta}_{\mathcal{A}^{0c}})). \tag{14}$$

Consider the following parameter set $\mathcal{W}_n \equiv \{\boldsymbol{\beta}_n \in \mathcal{V}_{\alpha_n}(\boldsymbol{\beta}_n^0); \|\boldsymbol{\beta}_{\mathcal{A}^{0c}}\|_2 > 0\}$. We show that $\mathbb{P}[\widehat{\boldsymbol{\beta}}_n \in \mathcal{W}_n] \to 0$, as $n \to \infty$. Let $\boldsymbol{\beta}_n = (\boldsymbol{\beta}_{\mathcal{A}^0}, \boldsymbol{\beta}_{\mathcal{A}^{0c}}) \in \mathcal{W}_n$ and an another parameter $\widetilde{\boldsymbol{\beta}}_n = (\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^0}, \widetilde{\boldsymbol{\beta}}_{\mathcal{A}^{0c}}) \in \mathcal{V}_{\alpha_n}(\boldsymbol{\beta}_n^0)$, such that $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^0} = \boldsymbol{\beta}_{\mathcal{A}^0}$ and $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^{0c}} = \mathbf{0}$. Define

$$D_n(\boldsymbol{\beta}_n, \widetilde{\boldsymbol{\beta}}_n) \equiv Q_n(\boldsymbol{\beta}_n) - Q_n(\widetilde{\boldsymbol{\beta}}_n) = \frac{1}{2n} \sum_{i=1}^{n} \left[\rho_\tau(Y_i - \mathbf{X}_i^t \boldsymbol{\beta}_n) - \rho_\tau(Y_i - \mathbf{X}_i^t \widetilde{\boldsymbol{\beta}}_n) \right] + \sum_{j \in \mathcal{A}^{0c}} p_{SELO}(\beta_j). \tag{15}$$

Concerning the penalty of relation (15), as in the proof of Lemma A.2 of [Dicker et al. (2013)], relation (A.7), we have that there exists $C > 0$ such that

$$\sum_{j \in \mathcal{A}^{0c}} p_{SELO}(\beta_j) \geq \frac{\lambda_n}{\log 2} \log \left(\frac{C}{C + \gamma_n \alpha_n} + 1 \right) \|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2.$$

On the other hand, by assumption (A3), we have that there exists $C_1 > 0$ such that $\liminf_{n \to \infty} \left(\log \left(C/(C + \gamma_n \alpha_n) + 1 \right) \right) > C_1 > 0$. Then, for n large enough, there exists $\tilde{C} > 0$ such that

$$\sum_{j \in \mathcal{A}^{0c}} \frac{p_{SELO}(\beta_j)}{\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2} \geq \tilde{C} \lambda_n. \tag{16}$$

Let be the identity that follows from [Knight (1998)], for any $x, y \in \mathbb{R}$,

$$\rho_\tau(x - y) - \rho_\tau(x) = y(\mathbb{1}_{x \leq 0} - \tau) + \int_0^y (\mathbb{1}_{x \leq t} - \mathbb{1}_{x \leq 0}) dt.$$

Using this relation for the first sum of (15), we obtain:

$$\frac{1}{2n} \sum_{i=1}^{n} \left[\rho_\tau(Y_i - \mathbf{X}_i^t \boldsymbol{\beta}_n) - \rho_\tau(Y_i - \mathbf{X}_i^t \widetilde{\boldsymbol{\beta}}_n) \right] = \frac{1}{2n} (\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n)^t \sum_{i=1}^{n} \mathbf{X}_i [\mathbb{1}_{Y_i - \mathbf{X}_i^t \widetilde{\boldsymbol{\beta}}_n \leq 0} - \tau]$$
$$+ \frac{1}{n} \sum_{i=1}^{n} \int_0^{\mathbf{X}_i^t(\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n)} [\mathbb{1}_{Y_i - \mathbf{X}_i^t \widetilde{\boldsymbol{\beta}}_n \leq t} - \mathbb{1}_{Y_i - \mathbf{X}_i^t \widetilde{\boldsymbol{\beta}}_n \leq 0}] dt \equiv T_{1n} + T_{2n} \tag{17}$$

For T_{1n} we have, by assumption (A3) and since the density f is bounded in a neighborhood of 0:

$$\mathbb{E}[T_{1n}] = (\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n)^t \frac{1}{2n} \sum_{i=1}^{n} \mathbf{X}_i [F(\mathbf{X}_i^t(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0)) - F(0)] = (\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n)^t \frac{1}{2n} \left(\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^t \right) (\boldsymbol{\beta}_n^0 - \widetilde{\boldsymbol{\beta}}_n) f(0)(1 + o(1)).$$

Then $|\mathbb{E}[T_{1n}]| \leq \|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2 \|(2n)^{-1} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^t\|_2 \|\boldsymbol{\beta}_n^0 - \widetilde{\boldsymbol{\beta}}_n\|_2 f(0)(1 + o(1))$. Since the matrix $n^{-1} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^t$ is Hermitian, we have that $\|n^{-1} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^t\|_2 = \lambda_{\max}(n^{-1} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^t) \leq R_0$. Hence, by (A2), we have $|\mathbb{E}[T_{1n}]| \leq \|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2 \|\boldsymbol{\beta}_n^0 - \widetilde{\boldsymbol{\beta}}_n\|_2 R_0 f(0)$. Therefore $\mathbb{E}[T_{1n}] = O(\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2 \|\boldsymbol{\beta}_n^0 - \widetilde{\boldsymbol{\beta}}_n\|_2) = O(\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2^2)$. By calculations analogous to $\mathbb{E}[T_{1n}]$, using independence of ε_i, we have that $\mathbb{E}[T_{1n}^2] = Cn^{-1} \|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|^3 \to 0$, for $n \to \infty$. Since $Var[T_{1n}] \leq \mathbb{E}[T_{1n}^2]$, using Bienaymé-Tchebychev inequality, we obtain

$$T_{1n} = C \|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2^2 (1 + o_{\mathbb{P}}(1)). \tag{18}$$

Study now T_{2n} of (17), which can be written as: $T_{2n} = n^{-1} \sum_{i=1}^{n} \int_0^{\mathbf{X}_i^t(\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n)} [\mathbb{1}_{\varepsilon_i \leq t - \mathbf{X}_i^t(\boldsymbol{\beta}_n^0 - \widetilde{\boldsymbol{\beta}}_n)} - \mathbb{1}_{\varepsilon_i \leq -\mathbf{X}_i^t(\boldsymbol{\beta}_n^0 - \widetilde{\boldsymbol{\beta}}_n)}] dt$. Then, taking into account that $\boldsymbol{\beta}_n \in \mathcal{V}_{\alpha_n}(\boldsymbol{\beta}_n^0)$, together with assumptions (A1), (A3), we have

$$\mathbb{E}[T_{2n}] = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\mathbf{X}_i^t(\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n)} [F(t - \mathbf{X}_i^t(\boldsymbol{\beta}_n^0 - \widetilde{\boldsymbol{\beta}}_n)) - F(-\mathbf{X}_i^t(\boldsymbol{\beta}_n^0 - \widetilde{\boldsymbol{\beta}}_n))] dt = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\mathbf{X}_i^t(\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n)} [t \cdot f(\mathbf{X}_i^t(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0)) + o(t)] dt.$$

By Theorem 1, together with assumptions (A1), (A3), we have that $f(\mathbf{X}_i^t(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0))$ is bounded by a constant $C \in (0, \infty)$. Thus, as for T_{1n}, using assumption (A2) and the fact that $n^{-1} \sum_{i=1}^{n} \|\mathbf{X}_i\|_2^2 - tr(n^{-1} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^t) \to 0$,

we have $|E[T_{2n}]| \leq Cn^{-1}\sum_{i=1}^{n}\|\mathbf{X}_i\|_2^2\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^0\|_2 + o(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i^t(\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n)) = C\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2^2$. We show similarly that $E[T_{2n}^2] = Cn^{-1}\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|^3$. Then, by Bienaymé-Tchebychev inequality, we have:

$$T_{2n} = C\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2^2(1 + o_{\mathbb{P}}(1)). \tag{19}$$

Hence, by relations (18), (19), we obtain

$$T_{1n} + T_{2n} = C\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2^2(1 + o_{\mathbb{P}}(1)).$$

Thus, taking into account this last relation together with relations (15), (16), (17), and since $\boldsymbol{\beta}_n \in \mathcal{W}_n$, we have: $D_n(\boldsymbol{\beta}_n, \widetilde{\boldsymbol{\beta}}_n)\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2^{-1} \geq C\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2 + \widetilde{C}\lambda_n$. Since $\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\| = O(\alpha_n)$ and $\lambda_n/\alpha_n \to \infty$ by (A5), we have that there exists $C_+ > 0$ such that $D_n(\boldsymbol{\beta}_n, \widetilde{\boldsymbol{\beta}}_n)\|\boldsymbol{\beta}_n - \widetilde{\boldsymbol{\beta}}_n\|_2^{-1} > C_+\lambda_n > 0$. But for $\boldsymbol{\beta}_n^0$, taking into account the definition of $\widetilde{\boldsymbol{\beta}}_n$ and that of $D_n(\boldsymbol{\beta}_n^0, \widetilde{\boldsymbol{\beta}}_n) \equiv Q_n(\boldsymbol{\beta}_n^0) - Q_n(\widetilde{\boldsymbol{\beta}}_n)$, we have that $D_n(\boldsymbol{\beta}_n^0, \widetilde{\boldsymbol{\beta}}_n) = C\|\boldsymbol{\beta}_n^0 - \widetilde{\boldsymbol{\beta}}_n\|_2^2(1 + o_{\mathbb{P}}(1))$. Then, by (A3), we have $\mathbb{P}[\boldsymbol{\beta}_n \in \mathcal{W}_n] \to 0$ and relation (14) follows.

(ii) Taking into account the estimator convergence rate obtained by Theorem 1 and claim *(i)*, the estimator $\widehat{\boldsymbol{\beta}}_n$ can be written $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_n^0 + \alpha_n\boldsymbol{\delta}$, with, $\boldsymbol{\delta} \equiv (\delta_1, \cdots, \delta_{d_n}) \in \mathbb{R}^{d_n}$, $\boldsymbol{\delta}_{\mathcal{A}^{0c}} = \mathbf{0}$ and $\|\boldsymbol{\delta}_{\mathcal{A}^0}\|_2^2 \leq C|\mathcal{A}^0|$. Consider then

$$Q_n(\boldsymbol{\beta}_n^0 + \alpha_n\boldsymbol{\delta}) - Q_n(\boldsymbol{\beta}_n^0) = \frac{1}{2n}\sum_{i=1}^{n}[\rho_\tau(Y_i - \mathbf{X}_i^t(\boldsymbol{\beta}_n^0 + \alpha_n\boldsymbol{\delta})) - \rho_\tau(\varepsilon_i)] + \mathcal{P}, \tag{20}$$

with $\mathcal{P} \equiv \sum_{j\in\mathcal{A}^0} p_{SELO}(\beta_j^0 + \alpha_n\delta_j) - p_{SELO}(\beta_j^0)$. Let us first study $\mathcal{P}$. For any $j \in \mathcal{A}^0$, by Lemma 1, we have that there exists a constant $\widetilde{C}_j$ such that

$$p_{SELO}(\beta_j^0 + \alpha_n\delta_j) - p_{SELO}(\beta_j^0) = \frac{\lambda_n}{\log 2}[g(\beta_j^0 + \alpha_n\delta_j) - g(\beta_j^0)] = \frac{\lambda_n}{\log 2}\gamma_n\left(|\beta_j^0 + \alpha_n\delta_j| - |\beta_j^0|\right)\widetilde{C}_j,$$

with $|\widetilde{C}_j| < \infty$, for any $j \in \mathcal{A}^0$. Since $\alpha_n \to 0$, $|\beta_j^0| > C > 0$, $\forall j \in \mathcal{A}^0$ and δ_j bounded, we have that for n large enough, the parameters $\beta_j^0 + \alpha_n\delta_j$ and β_j^0 have the same sign. Then

$$\mathcal{P} = C\frac{\lambda_n}{\log 2}\gamma_n\alpha_n\sum_{j\in\mathcal{A}^0}(\pm\delta_j) = C\lambda_n\alpha_n\gamma_n|\mathcal{A}^0|. \tag{21}$$

For the first term of the right-hand side of (20) we have:

$$\frac{1}{2n}\sum_{i=1}^{n}[\rho_\tau(\varepsilon_i - \alpha_n\mathbf{X}_i^t\boldsymbol{\delta}) - \rho_\tau(\varepsilon_i)] = \frac{\alpha_n}{2n}\sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\delta}[\mathbb{1}_{\varepsilon_i\leq 0} - \tau] + \frac{1}{2n}\sum_{i=1}^{n}\int^{\alpha_n\mathbf{X}_i^t\boldsymbol{\delta}}[\mathbb{1}_{\varepsilon_i\leq t} - \mathbb{1}_{\varepsilon_i\leq 0}]dt \equiv J_1 + J_2.$$

Since $E[J_1] = 0$, using independence of (ε_i), assumption (A4) and the Cauchy-Schwarz inequality, we get that

$$Var(J_1) \leq E[J_1^2] = \frac{\alpha_n^2}{4n^2}\tau(1-\tau)\sum_{i=1}^{n}(\mathbf{X}_i^t\boldsymbol{\delta})^2 \leq \frac{\alpha_n^2}{4n^2}\tau(1-\tau)\sum_{i=1}^{n}\|\mathbf{X}_{i,\mathcal{A}^0}^t\|_2^2\|\boldsymbol{\delta}_{\mathcal{A}^0}\|_2^2 = \frac{C}{n}\alpha_n^2|\mathcal{A}^0| \leq \alpha_n^2\frac{d_n}{n} \to 0.$$

For J_2 we have:

$$E[J_2] = \frac{1}{2n}\sum_{i=1}^{n}\int_0^{\alpha_n\mathbf{X}_i^t\boldsymbol{\delta}}\left(tf(0) + o(t^2)\right)dt = \frac{1}{4}f(0)\alpha_n^2\boldsymbol{\delta}^t\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^t\right)\boldsymbol{\delta}(1 + o(1)). \tag{22}$$

Using assumption (A2), we have that $f(0)\alpha_n^2\|\boldsymbol{\delta}\|_2^2\cdot\lambda_{\min}\left(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^t\right) \leq E[J_2] \leq f(0)\alpha_n^2\|\boldsymbol{\delta}\|_2^2\cdot\lambda_{\max}\left(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^t\right)$. Taking into account the fact that $\|\boldsymbol{\delta}\|_2^2 = \|\boldsymbol{\delta}_{\mathcal{A}^0}\|_2^2 \leq C|\mathcal{A}^0|$, we have $E[J_2] = Cf(0)\alpha_n^2|\mathcal{A}^0|$. We prove similarly $Var(J_2) = O(n^{-1}\alpha_n^3|\mathcal{A}^0|)$. We compare $\alpha_n^2|\mathcal{A}^0|$ with $\lambda_n\alpha_n\gamma_n|\mathcal{A}^0|$ obtained by (21) for the penalty, $\frac{\alpha_n^2|\mathcal{A}^0|}{\lambda_n\alpha_n\gamma_n|\mathcal{A}^0|} = \frac{\alpha_n}{\lambda_n\gamma_n}$. By (A5), $\gamma_n = O\left(\frac{\alpha_n^3}{d_n}\right)$, thus $\frac{\gamma_n}{\alpha_n} = \frac{\alpha_n^2}{d_n}$. Then $\frac{\alpha_n}{\lambda_n\gamma_n} = \frac{d_n}{\alpha_n^2}\frac{1}{\lambda_n} = \frac{n}{\lambda_n} \to \infty$, as $n \to \infty$. Thus, minimizing (20) amounts to minimizing $J_1 + J_2$, with respect to $\alpha_n\boldsymbol{\delta}$. Using (22), we obtain:

$$\frac{1}{2n}\sum_{i=1}^{n}[\rho_\tau(Y_i - \mathbf{X}_i^t(\boldsymbol{\beta}_n^0 + \alpha_n\boldsymbol{\delta})) - \rho_\tau(\varepsilon_i)] = \frac{\alpha_n}{2n}\sum_{i=1}^{n}\mathbf{X}_{i,\mathcal{A}^0}^t\boldsymbol{\delta}_{\mathcal{A}^0}[\mathbb{1}_{\varepsilon_i<0} - \tau] + \frac{1}{4}f(0)\alpha_n^2\boldsymbol{\delta}_{\mathcal{A}^0}^t\boldsymbol{\Sigma}_{\mathcal{A}^0}\boldsymbol{\delta}_{\mathcal{A}^0}(1 + o_{\mathbb{P}}(1)). \tag{23}$$

The minimizer of (23) is:

$$\alpha_n \boldsymbol{\delta}_{\mathcal{A}^0} = -\frac{1}{n}\frac{1}{f(0)}\boldsymbol{\Sigma}_{\mathcal{A}^0}^{-1}\Big(\sum_{i=1}^n \mathbf{X}_{i,\mathcal{A}^0}(\mathbb{1}_{\varepsilon_i \le 0} - \tau)\Big). \tag{24}$$

For studying (24), let us consider the following independent random variable sequence $W_i \equiv (f(0))^{-1}\mathbf{u}^t\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\mathbf{X}_{i\mathcal{A}}(\mathbb{1}_{\varepsilon_i \le 0} - \tau)$, with $\mathbf{u}$ a vector of dimension $|\mathcal{A}^0|$ and such that $\|\mathbf{u}\|_2 = 1$. We have that $I\!\!E[W_i] = 0$ and $\sum_{i=1}^n Var(W_i) = n\tau(1-\tau)(f(0))^{-2}\mathbf{u}^t\boldsymbol{\Sigma}_{\mathcal{A}^0}^{-1}\mathbf{u}$. Then, by CLT for independent random variable sequences (W_i), we have

$$\sqrt{n}f(0)\frac{\mathbf{u}^t(\widehat{\boldsymbol{\beta}}_{\mathcal{A}^0} - \boldsymbol{\beta}_{\mathcal{A}^0}^0)}{\sqrt{\tau(1-\tau)(\mathbf{u}^t\boldsymbol{\Sigma}_{\mathcal{A}^0}^{-1}\mathbf{u})}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1). \tag{25}$$

Claim (ii) results taking into account the fact that $\widehat{\boldsymbol{\beta}}_{\mathcal{A}^0} - \boldsymbol{\beta}_{\mathcal{A}^0}^0 = \alpha_n \boldsymbol{\delta}_{\mathcal{A}^0}$ and relations (24), (25). ■

Remark 2 *The cardinal of the set $\mathcal{A}^0$ may depend on n and converge to ∞ as $n \to \infty$.*

3   Tuning parameter selection

In this section we propose a criterion of type BIC to select the tuning parameters λ and γ. This criterion will also estimate the set $\mathcal{A}^0$, defined by (13). We start with introducing some notations.
- $\mathcal{A}_n$ a some index set $\subseteq \{1, \cdots, d_n\}$, which does not depend on tuning parameters.
- $(\lambda, \gamma) \in (0, \infty)^2$ some tuning parameters, which does not depend on n.
- $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)$ the seamless L_0 quantile estimator of $\boldsymbol{\beta}_{\mathcal{A}_n}$ obtained on some index set $\mathcal{A}_n \subset \{1, \cdots, d_n\}$ and with λ, γ as tuning parameters. We denote its components by $\widehat{\beta}_{\mathcal{A}_n,j}(\lambda, \gamma)$, for $j \in \mathcal{A}_n$.
- $\widehat{\boldsymbol{\beta}}(\lambda_n, \gamma_n)$ the seamless L_0 quantile estimator of $\boldsymbol{\beta}$ obtained on the index set $\{1, \cdots, d_n\}$, with (λ_n, γ_n) as tuning parameters. Then $\widehat{\boldsymbol{\beta}}(\lambda_n, \gamma_n) = \widehat{\boldsymbol{\beta}}_n$, with $\widehat{\boldsymbol{\beta}}_n$ obtained by (4). We denote its components by $\widehat{\beta}_j(\lambda_n, \gamma_n)$, for $j \in \{1, \cdots, d_n\}$.
- $\widehat{\mathcal{A}}_{\widehat{\boldsymbol{\beta}}(\lambda_n, \gamma_n)} \equiv \{j \in \{1, \cdots, d_n\}; \quad \widehat{\beta}_j(\lambda_n, \gamma_n) \ne 0\}$.
- (λ_n, γ_n) is a tuning parameter sequence such that: $\lim_{n\to\infty} \mathbb{P}[\widehat{\mathcal{A}}_{\widehat{\boldsymbol{\beta}}(\lambda_n, \gamma_n)} = \mathcal{A}^0] = 1$.

In order to define the BIC criterion, let us consider $(S_n)_{n\ge 1}$, a sequence of real numbers, defined as:

- if $d_n/\log n = o(1)$, we consider $S_n = 1$ for any $n \in \mathbb{N}$;

- if $d_n/\log n \ne o(1)$, we consider (S_n) a sequence converging to ∞ such that $\dfrac{d_n}{S_n \log n} \to 0$, $\dfrac{\log n}{n}|\mathcal{A}^0|S_n \to 0$.

In order to select $\mathcal{A}_n$, λ et γ, we consider the following BIC criterion:

$$BIC(\mathcal{A}_n; (\lambda, \gamma)) \equiv \log\Big(\frac{1}{n}\sum_{i=1}^n \rho(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma))\Big) + \frac{\log n}{n}S_n\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)\|_0, \tag{26}$$

with $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)\|_0 = \sum_{j=1}^{d_n}\mathbb{1}_{\widehat{\beta}_{\mathcal{A}_n,j}(\lambda, \gamma)\ne 0}$. For the tuning parameters λ_n, γ_n and the estimator $\widehat{\boldsymbol{\beta}}(\lambda_n, \gamma_n)$, let us consider the value of the BIC criterion corresponding to (26):

$$BIC(\lambda_n, \gamma_n) \equiv \log\Big(\frac{1}{n}\sum_{i=1}^n \rho_\tau\big(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n, \gamma_n)\big)\Big) + \frac{\log n}{n}S_n\|\widehat{\boldsymbol{\beta}}(\lambda_n, \gamma_n)\|_0.$$

If the conditions of Theorem 2 are satisfied, then $\widehat{\boldsymbol{\beta}}(\lambda_n, \gamma_n)$ satisfies the sparsity property,

$$\lim_{n\to\infty} \mathbb{P}[\{j \in \{1, \cdots, d_n\}; \widehat{\beta}_j(\lambda_n, \gamma_n) \ne 0\} = \mathcal{A}^0] = 1.$$

In order to prove, by the following theorem, that the BIC criterion selects correctly, with a probability converging to 1, the tuning parameters λ and γ, we will consider the index sets $\mathcal{A}_n$ such that $|\mathcal{A}_n| \le s_n$, with the assumption $s_n = O(n^a)$, $0 < a < 1/2$. Consider also two index sets A_{1n} et A_{2n}:

$$A_{1n} \equiv \{\mathcal{A}_n; \mathcal{A}^0 \subset \mathcal{A}_n, \mathcal{A}^0 \ne \mathcal{A}_n, |\mathcal{A}_n| \le s_n\}, \qquad A_{2n} \equiv \{\mathcal{A}_n; \mathcal{A}^0 \not\subseteq \mathcal{A}_n, |\mathcal{A}_n| \le s_n\}.$$

Theorem 3 *We suppose that $0 < I\!E[\rho_\tau(\varepsilon)] < \infty$. Then, if instead of assumption (A4) we take $n^{(a-1)/2}d_n^{1/2} \to 0$ as $n \to \infty$, under (A1)-(A3), (A5), we have:*

$$\lim_{n\to\infty} \mathbb{P}\Big[\min_{\substack{\mathcal{A}_n\subseteq\{1,\cdots,d_n\},(\lambda,\gamma)\in(0,\infty)^2\\|\mathcal{A}_n|\leq s_n}} BIC(\mathcal{A}_n;(\lambda,\gamma)) = BIC(\lambda_n,\gamma_n)\Big] = 1.$$

Proof. The theorem is proved if the following two statements are shown:

$$\lim_{n\to\infty} \mathbb{P}\Big[\min_{\mathcal{A}_n\in A_{1n}} BIC(\mathcal{A}_n;(\lambda,\gamma)) > BIC(\lambda_n,\gamma_n)\Big] = 1, \tag{27}$$

$$\lim_{n\to\infty} \mathbb{P}\Big[\min_{\mathcal{A}_n\in A_{2n}} BIC(\mathcal{A}_n;(\lambda,\gamma)) > BIC(\lambda_n,\gamma_n)\Big] = 1. \tag{28}$$

Proof of relation (27). Since $\mathcal{A}_n \in A_{1n}$, then $|\mathcal{A}_n| > |\mathcal{A}^0|$. Let us consider the difference

$$\begin{aligned}
&BIC(\mathcal{A}_n;(\lambda,\gamma)) - BIC(\lambda_n,\gamma_n)\\
&= \log\Big(1 + \frac{n^{-1}\rho_\tau(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda,\gamma)) - n^{-1}\sum_{i=1}^n\rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n))}{n^{-1}\sum_{i=1}^n\rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n))}\Big) + \frac{\log n}{n}S_n\big[|\mathcal{A}_n| - |\mathcal{A}^0|\big].
\end{aligned}$$

In addition of index set $\mathcal{A}_n \in A_{1n}$, let us consider the following sets: $\widehat{\mathcal{A}}_1 = \{j; \widehat{\beta}_{n,j} \neq 0\}$ and $\widehat{\mathcal{A}}_2 = \{j; \widehat{\beta}_{\mathcal{A}_n,j}(\lambda,\gamma) \neq 0\}$. Recall that $\widehat{\boldsymbol{\beta}}_n$ is $\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n)$. Since $\mathcal{A}^0 \subset \mathcal{A}_n$, by Theorem 2(i), we have that, $\lim_{n\to\infty}\mathbb{P}\big[\widehat{\mathcal{A}}_1 = \widehat{\mathcal{A}}_2 = \mathcal{A}^0\big] = 1$. Without loss of generality, we suppose that $\mathcal{A}^0 \subseteq \widehat{\mathcal{A}}_1 \subseteq \widehat{\mathcal{A}}_2$, the other cases are similar. Using the elementary inequality $|\rho_\tau(u-v) - \rho_\tau(u)| < |v|$, for all $u,v \in \mathbb{R}$, we have, with probability one,

$$\begin{aligned}
n^{-1}\Big|\sum_{i=1}^n\big[\rho_\tau(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda,\gamma)) - \rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n))\big]\Big| &\leq n^{-1}\sum_{i=1}^n\Big|\mathbf{X}_{i,\widehat{\mathcal{A}}_2}^t\big(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda,\gamma) - \widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n)\big)_{\widehat{\mathcal{A}}_2}\Big|\\
&\leq \max_{1\leq i\leq n}\|\mathbf{X}_{i,\widehat{\mathcal{A}}_2}^t\|_2\Big\|\big(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda,\gamma) - \widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n)\big)_{\widehat{\mathcal{A}}_2}\Big\|_2
\end{aligned}$$

which is, by assumption (A3) and Theorem 1, $o(\alpha_n)O_\mathbb{P}(\alpha_n) = o_\mathbb{P}(1)$. For the second inequality, the estimators $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda,\gamma)$ were completed by with zeros for obtaining a vector of dimension d_n. Then

$$n^{-1}\sum_{i=1}^n\rho_\tau(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda,\gamma)) - n^{-1}\sum_{i=1}^n\rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n)) \xrightarrow[n\to\infty]{\mathbb{P}} 0. \tag{29}$$

In the same way, we have: $n^{-1}\sum_{i=1}^n\big(\rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n)) - \rho_\tau(\varepsilon_i)\big) \xrightarrow[n\to\infty]{\mathbb{P}} 0$. On the other hand, be the LLN,

$$\frac{1}{n}\sum_{i=1}^n\rho_\tau(\varepsilon_i) \xrightarrow[n\to\infty]{\mathbb{P}} I\!E[\rho_\tau(\varepsilon)] \in (0,\infty). \tag{30}$$

Taking into account (29) and (30), we can apply the inequality $\log(1+x) \geq -2|x|$ for all $|x| < 1/2$,

$$\begin{aligned}
&\log\Big(1 + \frac{n^{-1}\rho_\tau(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda,\gamma)) - n^{-1}\sum_{i=1}^n\rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n))}{n^{-1}\sum_{i=1}^n\rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n))}\Big)\\
&\geq -2\frac{\big|n^{-1}\rho_\tau(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda,\gamma)) - n^{-1}\sum_{i=1}^n\rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n))\big|}{n^{-1}\sum_{i=1}^n\rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n))}
\end{aligned} \tag{31}$$

But, by the proof of Theorem 1, relation (11), we have, with probability tending to 1: $n^{-1}\rho_\tau(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda,\gamma)) - n^{-1}\sum_{i=1}^n\rho_\tau(Y_i - \mathbf{X}_i^t\widehat{\boldsymbol{\beta}}(\lambda_n,\gamma_n)) = C\alpha_n^2$, with $C > 0$, for n large enough. Using (31), we have:

$$\min_{\mathcal{A}_n\in A_{1n}}\big(BIC(\mathcal{A}_n;(\lambda,\gamma)) - BIC(\lambda_n,\gamma_n)\big) \geq \min_{\mathcal{A}_n\in A_{1n}}\Big(-\tilde{C}\frac{d_n}{n}|\mathcal{A}_n| + \frac{\log n}{n}C_n\big(|\mathcal{A}_n| - |\mathcal{A}^0|\big)\Big) > C > 0,$$

8

with $\tilde{C} > 0$. So, relation (27) is proved.

$\underline{\textit{Proof relation (28)}}$. Let be the index sets $\mathcal{A}_n \in A_{2n}$ and $\widetilde{\mathcal{A}}_n \equiv \mathcal{A}_n \cup \mathcal{A}^0$.
Let $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)$ be the estimator of dimension $|\mathcal{A}_n|$ built on the variables $\mathbf{X}_{\mathcal{A}_n}$. Let also $\widetilde{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n}$ equal to $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)$ on $\mathcal{A}_n$ and completed with 0 to obtain a vector of dimension $|\widetilde{\mathcal{A}}_n|$. Then, denoting $b^0 \equiv \min_{j \in \mathcal{A}^0} |\beta_j^0| > 0$, we have

$$\|\widetilde{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n} - \boldsymbol{\beta}_{\widetilde{\mathcal{A}}_n}^0\|_2 > \|\boldsymbol{\beta}_{\mathcal{A}^0 \setminus \mathcal{A}_n}^0\|_2 \geq b^0.$$

Then, since $\rho_\tau(.)$ is convex, we have that there exists $\bar{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n} \in \mathbb{R}^{|\widetilde{\mathcal{A}}_n|}$, with $\|\bar{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n} - \boldsymbol{\beta}_{\widetilde{\mathcal{A}}_n}^0\|_2 = b^0$, such that $\sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)) \geq \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{i,\widetilde{\mathcal{A}}_n}^t \bar{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n})$. Thus

$$\begin{aligned}
\mathcal{R} &\equiv n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)) - \frac{1}{n} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{i,\widetilde{\mathcal{A}}_n}^t \widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n}) \\
&\geq n^{-1} \Big[\inf_{(\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n} \in \mathcal{B}_\delta(\widetilde{\mathcal{A}}_n)} \mathbb{E}\big[G_{n,\widetilde{\mathcal{A}}_n}((\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n}) \big] - \sup_{(\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n} \in \mathcal{B}_\delta(\widetilde{\mathcal{A}}_n)} \big| G_{n,\widetilde{\mathcal{A}}_n}((\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n}) \\
&\quad - \mathbb{E}\big[G_{n,\widetilde{\mathcal{A}}_n}((\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n}) \big] \big| - G_{n,\widetilde{\mathcal{A}}_n}(\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n} - \boldsymbol{\beta}_{\widetilde{\mathcal{A}}_n}^0) \Big],
\end{aligned}$$

with $G_{n,\widetilde{\mathcal{A}}_n}(\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n} - \boldsymbol{\beta}_{\widetilde{\mathcal{A}}_n}^0) \equiv \sum_{i=1}^n \big[\rho_\tau(Y_i - \mathbf{X}_{i,\widetilde{\mathcal{A}}_n}^t \widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n}) - \rho_\tau(\varepsilon_i) \big]$ and $G_{n,\widetilde{\mathcal{A}}_n}((\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n})$ defined similarly.
As for the calculation of relation (19), we have, with probability converging to 1:

$$\inf_{(\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n} \in \mathcal{B}_\delta(\widetilde{\mathcal{A}}_n)} \mathbb{E}\big[G_{n,\widetilde{\mathcal{A}}_n}((\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n}) \big] \geq C n (b^0)^2,$$

with $C > 0$ and

$$G_{n,\widetilde{\mathcal{A}}_n}(\widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n} - \boldsymbol{\beta}_{\widetilde{\mathcal{A}}_n}^0) = O_{\mathbb{P}}(d_n). \tag{32}$$

By Lemma 2 we have:

$$\sup_{(\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n} \in \mathcal{B}_\delta(\widetilde{\mathcal{A}}_n)} \big| G_{n,\widetilde{\mathcal{A}}_n}((\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n}) - \mathbb{E}\big[G_{n,\widetilde{\mathcal{A}}_n}((\boldsymbol{\beta} - \boldsymbol{\beta}^0)_{\widetilde{\mathcal{A}}_n}) \big] \big| = O_{\mathbb{P}}\big(d_n^{1/2} n^{(1+a)/2} \big).$$

Then, with probability converging to 1, as $n \to \infty$, we have $\mathcal{R} > C(b^0)^2 - d_n^{1/2} n^{(1+a)/2} n^{-1} - d_n n^{-1}$. Taking into account the assumption $n^{(a-1)/2} d_n^{1/2} \to 0$, we have that, for n large enough, with probability converging to 1,

$$\mathcal{R} > C(b^0)^2 > c_1 > 0. \tag{33}$$

Hence

$$\begin{aligned}
&\min_{\mathcal{A}_n \in A_{2n}} \big[BIC(\mathcal{A}_n; (\lambda, \gamma)) - BIC(\widetilde{\mathcal{A}}_n; (\lambda, \gamma)) \big] = \big(|\mathcal{A}_n| - |\widetilde{\mathcal{A}}_n| \big) \frac{\log n}{n} S_n \\
&+ \min_{\mathcal{A}_n \in A_{2n}} \log \Big(1 + \frac{n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{i,\mathcal{A}_n}^t \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)) - n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{i,\widetilde{\mathcal{A}}_n}^t \widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n})}{n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{i,\widetilde{\mathcal{A}}_n}^t \widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n})} \Big)
\end{aligned}$$

which is, with probability converging to 1, using (33):

$$\geq \min_{\mathcal{A}_n \in A_{2n}} \min \Big(\log 2, \frac{c_1}{n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{i,\widetilde{\mathcal{A}}_n}^t \widehat{\boldsymbol{\beta}}_{\widetilde{\mathcal{A}}_n})} \Big) - |\mathcal{A}^0| \frac{\log n}{n} S_n > 0. \tag{34}$$

The last inequality (> 0) results from (32) together with $\mathbb{E}[\rho_\tau(\varepsilon)] \in (0, \infty)$.
As for relation (27), we can prove, with probability tending to 1, for $n \to \infty$:

$$BIC(\widetilde{\mathcal{A}}_n; (\lambda, \gamma)) \geq \min_{\substack{\mathcal{A}_n' \\ \mathcal{A}_n' \in A_{1n}(2s_n)}} BIC(\mathcal{A}_n'; (\lambda, \gamma)) \geq BIC(\lambda_n, \gamma_n), \tag{35}$$

with $A_{1n}(2s_n) \equiv \{\mathcal{A}_n; \mathcal{A}^0 \subset \mathcal{A}_n, \mathcal{A}^0 \neq \mathcal{A}_n, |\mathcal{A}_n| \leq 2s_n\}$, $s_n = O(n^a)$, $a \in (0, 1/2)$. Then, with probability tending to 1, using (35) and (34), we have

$$
\begin{aligned}
&\min_{\mathcal{A}_n \in A_{2n}} BIC(\mathcal{A}_n; (\lambda, \gamma)) - BIC(\lambda_n, \gamma_n) \\
&= \min_{\mathcal{A}_n \in A_{2n}} \left[BIC(\mathcal{A}_n; (\lambda, \gamma)) - BIC(\widetilde{\mathcal{A}}_n; (\lambda, \gamma)) + BIC(\widetilde{\mathcal{A}}_n; (\lambda, \gamma)) - BIC(\mathcal{A}^0; (\lambda_n, \gamma_n)) \right] \\
&\geq \min_{\mathcal{A}_n \in A_{2n}} \left[BIC(\mathcal{A}_n; (\lambda, \gamma)) - BIC(\widetilde{\mathcal{A}}_n; (\lambda, \gamma)) \right] > 0
\end{aligned}
$$

and relation (28) is proved. ∎

Theorem 3 implies that we can choose as tuning parameters (λ_n, γ_n) thereby:

$$
(\widehat{\mathcal{A}}_n, \widehat{\lambda}_n, \widehat{\gamma}_n) \equiv \underset{\substack{\mathcal{A}_n \subseteq \{1, \cdots, d_n\}, (\lambda, \gamma) \in (0, \infty)^2 \\ |\mathcal{A}_n| \leq s_n}}{\arg\min} BIC(\mathcal{A}_n; (\lambda, \gamma)),
$$

choosing some s_n, such that $s_n = O(n^a)$, $a \in (0, 1/2)$. Obviously $\widehat{\mathcal{A}}_n = \mathcal{A}^0$ with a probability tending to 1. Then, in applications, we must first fix $\mathcal{A}_n, \lambda, \gamma$ and calculate:

$$
\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{A}_n|}}{\arg\min} Q_n(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{A}_n|}}{\arg\min} \left(\frac{1}{2n} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{X}_{i, \mathcal{A}_n}^t \boldsymbol{\beta}) + \frac{\lambda}{\log 2} \sum_{j \in \mathcal{A}_n} \log\left(\frac{|\beta_j|}{|\beta_j| + \gamma} + 1 \right) \right).
$$

Afterwards, we vary $\mathcal{A}_n, \lambda, \gamma$ on grid and take as $\widehat{\lambda}_n, \widehat{\gamma}_n$ and $\widehat{\mathcal{A}}_n = \widehat{\mathcal{A}}_{\widehat{\boldsymbol{\beta}}(\widehat{\lambda}_n, \widehat{\gamma}_n)}$:

$$
(\widehat{\mathcal{A}}_n, \widehat{\lambda}_n, \widehat{\gamma}_n) = \underset{\substack{\mathcal{A}_n \subseteq \{1, \cdots, d_n\}, (\lambda, \gamma) \in (0, \infty)^2 \\ |\mathcal{A}_n| \leq s_n}}{\arg\min} \left(\log\left(\frac{1}{n} \sum_{i=1}^n \rho_\tau\left(Y_i - \mathbf{X}_{i, \mathcal{A}_n}^t \widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)\right) \right) + \frac{\log n}{n} S_n \|\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n}(\lambda, \gamma)\|_0 \right).
$$

Then, we estimate simultaneously the best tuning parameters $\widehat{\lambda}_n$ and $\widehat{\gamma}_n$ and the parameters $\boldsymbol{\beta}$ that have components different of 0, such that the corresponding index set $\widehat{\mathcal{A}}_n$ is equal to $\mathcal{A}^0$, with probability tending to 1.

Remark 3 *Theorem 3 is the equivalent of Theorem 2 of [Li et al. (2012)], where the seamless L_0 penalized likelihood approach is considered, or of Theorem 2 of [Dicker et al. (2013)], for seamless L_0 penalized LS approach.*
In [Lee et al. (2014)], a BIC criterion is proposed to select the significant predictor variables of $\mathbf{X}$ in an high-dimensional quantile model.

Remark 4 *Algorithm and numerical part are a very difficult task, since $G_n(\boldsymbol{\beta}_n)$, defined by (5), is convex in $\boldsymbol{\beta}_n$ and the penalty $\mathcal{P}en(\boldsymbol{\beta}_n) = \sum_{j=1}^{d_n} p_{SELO}(\beta_j)$ is concave in $\boldsymbol{\beta}_n$, both being continuous, but not differentiable in $\boldsymbol{\beta}_n$. The same type of problem as ours, but with the process $G_n(\boldsymbol{\beta}_n)$ the likelihood (then differentiable in $\boldsymbol{\beta}_n$), was analyzed by [Dicker et al. (2013)]. They propose the coordinate descent algorithm to solve the optimization problem. For the method proposed in present paper, another work should be conducted on numerical method in order to find the seamless L_0 quantile estimator and the tuning parameters using the criterion given by Theorem 3.*

4 Lemmas

Lemma 1 *Let be the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(x) = \log(h(x) + 1)$, with the function $h : \mathbb{R} \to \mathbb{R}_+^*$, $h(x) = \dfrac{|x|}{|x| + \gamma_n}$. Then, $\forall x_1, x_2, C \in \mathbb{R}$ such that $|x_1|, |x_2| \geq C > 0$ and $|x_1| - |x_2| = o(1)$ we have that there exists $\tilde{C} > 0$ such that: $g(x_2) - g(x_1) = \tilde{C}\gamma_n(|x_2| - |x_1|)(-1)^{sgn(x_1 x_2)}$.*

Proof. By elementary calculus we have

$$
h(x_2) - h(x_1) = \gamma_n \frac{(|x_2| - |x_1|)(-1)^{sgn(x_1 x_2)}}{(|x_1| + \gamma_n)(|x_2| + \gamma_n)}.
$$

Then, taking into account the fact that for $|x| \simeq 0$ we have $\log(x+1) \simeq x$, the lemma follows. ∎

Let us consider the following notations:

$$
\begin{aligned}
\boldsymbol{\theta} &\equiv (\boldsymbol{\beta} - \boldsymbol{\beta^0})_{\mathcal{A}_n}, \\
g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}) &\equiv \rho_\tau(\varepsilon_i - \mathbf{X}_{i,\mathcal{A}_n}^t \boldsymbol{\theta}) - \rho_\tau(\varepsilon_i) - I\!\!E\left[\rho_\tau(\varepsilon_i - \mathbf{X}_{i,\mathcal{A}_n}^t \boldsymbol{\theta}) - \rho_\tau(\varepsilon_i)\right], \qquad \forall i = 1, \cdots, n, \\
\mathcal{B}_\delta(\mathcal{A}_n) &\equiv \{\boldsymbol{\theta} \in \mathbb{R}^{|\mathcal{A}_n|}; \|\boldsymbol{\theta}\|_2 \le \delta\}, \qquad \forall \delta > 0.
\end{aligned}
$$

Lemma 2 *Under assumptions (A2), (A3), if $s_n = O(n^a)$, with $a \in (0, 1/2)$, then, for any $\delta > 0$, we have*

$$
\sup_{\substack{\mathcal{A}_n \\ |\mathcal{A}_n| \le s_n}} \sup_{\boldsymbol{\theta} \in \mathcal{B}_\delta(\mathcal{A}_n)} \Big| \sum_{i=1}^n g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}) \Big| = O_\mathbb{P}\big(n^{(1+a)/2} d_n^{1/2}\big).
$$

Proof. The proof is similar to that of Lemma A.3 of [Lee et al. (2014)]. We consider for $k \ge 1$, $\Theta_n(2^{-k}\delta, \mathcal{A}_n)$ a grid of points in $\mathcal{B}_\delta(\mathcal{A}_n)$ such that for any $\boldsymbol{\theta} \in \mathcal{B}_\delta(\mathcal{A}_n)$ there exists $\boldsymbol{\theta}^{(k)} \in \Theta_n(2^{-k}\delta, \mathcal{A}_n)$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\|_2 \le \delta/2^k$. If we denote $M \equiv \max_{1 \le i \le d_n} \|\mathbf{X}_i\|_2$, then, for a given constant $C_1 > 0$, let we consider the natural number:

$$
K_n \equiv \min\big(k \ge 1; \frac{\delta}{2^k} \le \frac{C_1}{8M} n^{-1/2} d_n^{1/2}\big).
$$

Using the fact that for any $u, v \in \mathbb{R}$: $|\rho_\tau(u - v) - \rho_\tau(u)| < |v|$, then, we have with probability 1:

$$
\sup_{\boldsymbol{\theta} \in \mathcal{B}_\delta(\mathcal{A}_n)} \Big| \sum_{i=1}^n \big[g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}) - g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}^{(K_n)})\big]\Big| \le \frac{C_1}{2} n^{1/2} |\mathcal{A}_n|^{1/2} d_n^{1/2}. \tag{36}
$$

Denote

$$
\mathcal{P}_1 \equiv \mathbb{P}\Big[\sup_{\substack{\mathcal{A}_n \\ |\mathcal{A}_n| \le s_n}} \sup_{\boldsymbol{\theta} \in \mathcal{B}_\delta(\mathcal{A}_n)} \Big| \sum_{i=1}^n g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}) \Big| \ge C n^{(1+a)/2} d_n^{1/2}\Big]. \tag{37}
$$

Inequality (36) implies

$$
\mathcal{P}_1 \le \mathbb{P}\Big[\sup_{\substack{\mathcal{A}_n \\ |\mathcal{A}_n| \le s_n}} \sup_{\boldsymbol{\theta} \in \mathcal{B}_\delta(\mathcal{A}_n)} \Big| \sum_{i=1}^n g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}^{(K_n)}) \Big| \ge \frac{C}{2} n^{(1+a)/2} d_n^{1/2}\Big].
$$

On the other hand, for the cardinality $N_k(\mathcal{A}_n)$ of $\Theta_n(2^{-k}\delta, \mathcal{A}_n)$, we have: $N_k(\mathcal{A}_n) \le (1 + 4 \cdot 2^k)^{|\mathcal{A}_n|}$. Then

$$
\begin{aligned}
\mathcal{P}_1 &\le \sum_{\mathcal{A}_n, |\mathcal{A}_n| \le s_n} \mathbb{P}\Big[\sup_{\boldsymbol{\theta} \in \mathcal{B}_\delta(\mathcal{A}_n)} \sum_{k=1}^{K_n} |g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}^{(k)}) - g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}^{(k-1)})| > \frac{C}{2} n^{(1+a)/2} d_n^{1/2}\Big] \\
&\le \sum_{\mathcal{A}_n, |\mathcal{A}_n| \le s_n} \sum_{k=1}^{K_n} N_k(\mathcal{A}_n) N_{k-1}(\mathcal{A}_n) \max{}^* \mathbb{P}\Big[\Big| \sum_{i=1}^n [g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}^{(k)}) - g_{\mathcal{A}_n}(\varepsilon_i, \boldsymbol{\theta}^{(k-1)})]\Big| \ge \frac{C}{2} \eta_k n^{(1+a)/2} d_n^{1/2}\Big],
\end{aligned}
$$

with $\eta_k > 0$, $\sum_{k=1}^{K_n} \eta_k \le 1$. The $\max^*$ is calculated over all $\boldsymbol{\theta}^{(k)} \in \Theta_n(2^{-k}\delta, \mathcal{A}_n)$ and $\boldsymbol{\theta}^{(k-1)} \in \Theta_n(2^{-k+1}\delta, \mathcal{A}_n)$, with $\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}\|_2 \le 3 \cdot 2^{-k}\delta$. Moreover, by assumption (A2) we have: $n^{-1} \sum_{i=1}^n |\mathbf{X}_{i,\mathcal{A}_n}^t(\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)})|^2 \le 18 R_0 2^{-2k} \delta^2$. We take $\eta_k = \max\big(2^{-k} k^{1/2}/8, 96 R_0^{1/2} 2^{-k} s_n^{1/2} n^{-a/2} d_n^{-1/2} C^{-1} \log^{1/2}(1 + 4 \cdot 2^k)\big)$. By the Hoeffding inequality, we obtain:

$$
\begin{aligned}
\mathcal{P}_1 &\le \sum_{\mathcal{A}_n; |\mathcal{A}_n| \le s_n} \sum_{k=1}^{K_n} \exp\Big(2 s_n \log(1 + 4 \cdot 2^k) - \frac{C^2 \eta_k^2 n^a d_n}{48^2 \cdot R_0 \cdot 2^{-2k} \delta^2}\Big) \le 2 \sum_{\mathcal{A}_n; |\mathcal{A}_n| \le s_n} \sum_{k=1}^{K_n} \exp\Big(-\frac{C^2 k n^a d_n}{2 \cdot 8^2 \cdot 48^2 \cdot R_0 \cdot \delta^2}\Big) \\
&\le 4 \exp\Big(s_n \log d_n - \frac{C^2 k n^a d_n}{2 \cdot 8^2 \cdot 48^2 \cdot R_0 \cdot \delta^2}\Big) \to 0, \tag{38}
\end{aligned}
$$

the last relation following from the fact that $s_n = O(n^a)$. Lemma follows from relations (37) and (38). ∎

References

[Belloni and Chernozhukov (2011)] Belloni A., Chernozhukov V., (2011), l_1-Penalized quantile regression in high-dimensional sparse models. *Ann. Statist.*, **39**(1), 82-130.

[Ciuperca (2013)] Ciuperca G., (2013), Quantile regression in high-dimension with breaking, *J. Statist. Theory Appl.*, **12**(3), 288-305.

[Dicker et al. (2013)] Dicker L., Huang B., Lin X.,(2013). Variable selection and estimation with the seamless-L_0 penalty. *Statist. Sinica*, **23**, 929-962.

[Fan and Peng (2004)] Fan J., Peng H.,(2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*, **32**(3), 928-961.

[Fan et al. (2014a)] Fan J., Fan Y., Barut E.,(2014a). Adaptive robust variable selection. *Ann. Statist.*, **42**(1), 324-351.

[Fan et al. (2014b)] Fan J., Xue L., Zou H.,(2014b). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, **42**(3), 819-849.

[Knight (1998)] Knight K.,(1998). Limiting distributions for L1 regression estimators under general conditions. *Ann. Statist.*, **26**, 755-770.

[Lee et al. (2014)] Lee E.R., Noh B.H, Park B.U.,(2014). Model selection via Bayesian Information Criterion for quantile regression models. *J. Amer. Statist. Assoc.*, **109**, 216-229.

[Li et al. (2012)] Li Z., Wang S., Lin X.,(2012). Variable selection and estimation in generalized linear models with the seamless-L_0 penalty. *Canad. J. Statist.*, **40**(4), 745-769.

[Wu and Liu (2009)] Wu Y., Liu Y.,(2009). Variable selection in quantile regression. *Statist. Sinica*, **19**, 801-817.

[Wang et al. (2014)] Wang Z., Liu H., Zhang T., (2014) Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.*, **42**(6), 2164-2201.

[Zhang and Zhang (2012)] Zhang C.-H., Zhang T., (2012), A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.*, **27**(4), 576-593.

[Zheng et al.(2013)] Zheng Q., Gallagher C., Kulasekera K.B., (2013), Adaptive penalized quantile regression for high dimensional data. *J. Statist. Plann. Inference*, **143**(6), 1029-1038.

[Zou and Yuan(2008)] Zou H., Yuan M., (2008) Composite quantile regression and the oracle model selection theory. *Ann. Statist.*, **36**(3), 1108-1126.